Toward a Generic Subsetting Engine

Robert.O.Morris

Jet Propulsion Laboratory

California Institute of Technology

Abstract

By accepting a series of user-dictated constraints, software components we term *subsetters* can operate on extremely large data archives to yield a subset of the data that is more appropriate for the user's purposes. Compared to the original archives, subsetted data is smaller and more focused toward the user's needs. This makes it more manageable - easier to store, easier to transport and more relevant.

Data Engineers at the PO.DAAC are frequently required to develop this software and do so on a dataset-by-dataset basis. Each subsetter is essentially developed independently and from scratch. Any code sharing is strictly incidental and each Data Engineer has to develop their own independent conceptual model of the process by research and trial-and-error. The process is expensive and time consuming as a result.

We outline a plan by which we can construct a system that will subset a variety of datasets using the same model - a series of highly-generic reusable components that can be assembled appropriately to act in concert to subset any dataset. This common model will allow new Data Engineers to hurdle the subsetting learning curve more quickly and uniformly. They will have an existing model to follow and do not have to develop their own independent conceptual model. This common model will allow for shared learning and promote the use of a common vocabulary - desirable attributes when working in a team environment. Just as important, it speeds implementation through code reuse.